Warcbase:

Using Scalable Web Analytics to Analyze Canadian Collections En Masse

Ian Milligan
Assistant Professor
@ianmilligan1

Nick Ruest
Digital Assets Librarian
@ruebot

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

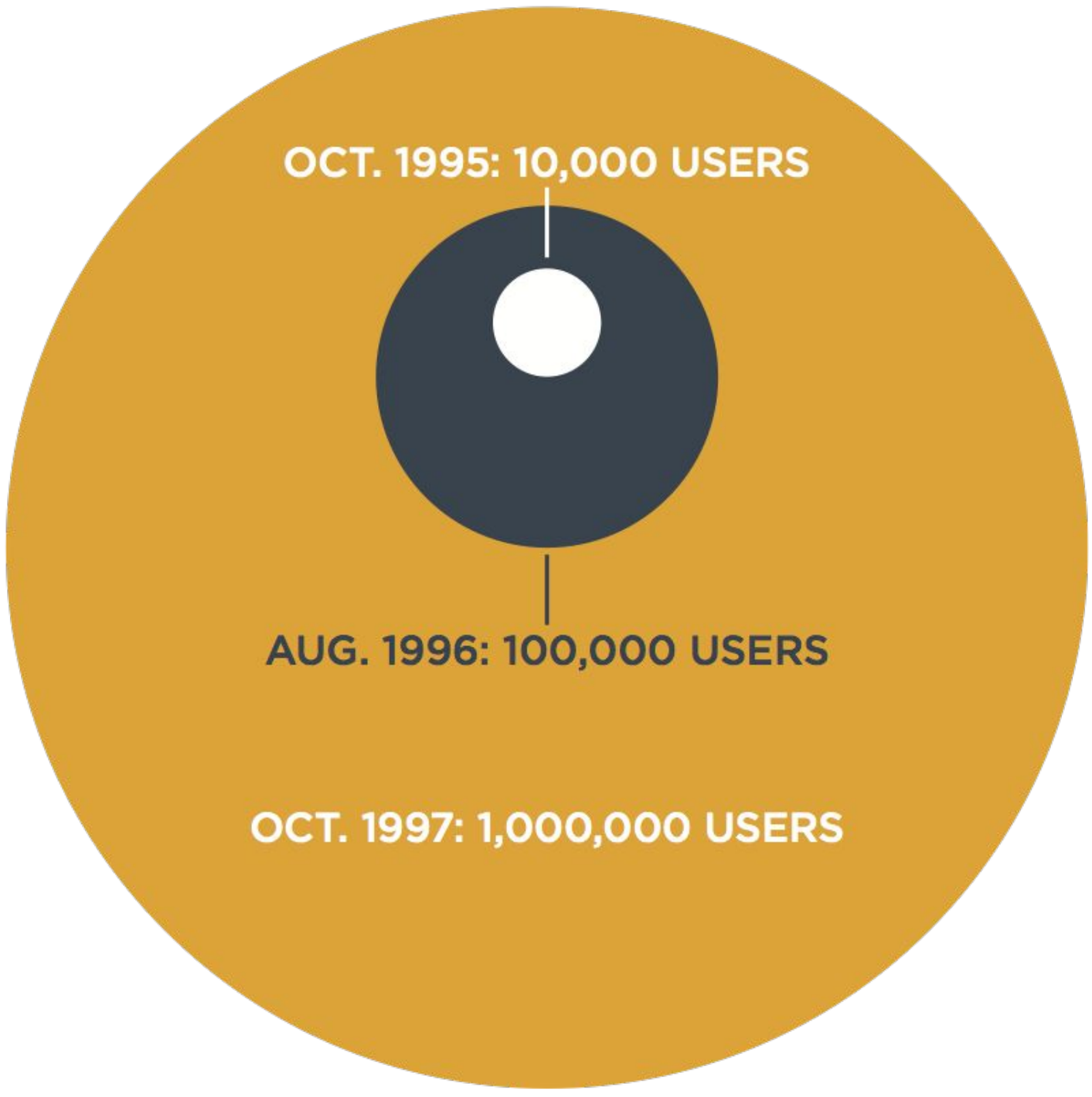
YORK
UNIVERSITÉ
UNIVERSITY

The Web as a Primary Source

- **Web archives will fundamentally affect the way historians write history**
  - We will have easier access to information on a previously-unknown scale, as well as improved capability to parse it;
  - This information will be left by people who rarely before left historical sources;

We live and experience our lives online, and historians need to study this.

And it’s all happening on a **scale historians have difficulty imagining!**




CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
PARTY LEADERSHIP / ADMINISTRATIVE FILES
Democratic Party
BOX 29
JOHN J. BURNS LIBRARY
BOSTON COLLEGE
374
CONGRESSIONAL ARCHIVES
THOMAS P. O'NEILL PAPERS
Series V
PARTY LEADERSHIP / ADMINISTRATIVE FILE
Democratic Party
BOX 30
JOHN J. BURNS LIBRARY
BOSTON COLLEGE

Scarcity

~~Scarcity~~

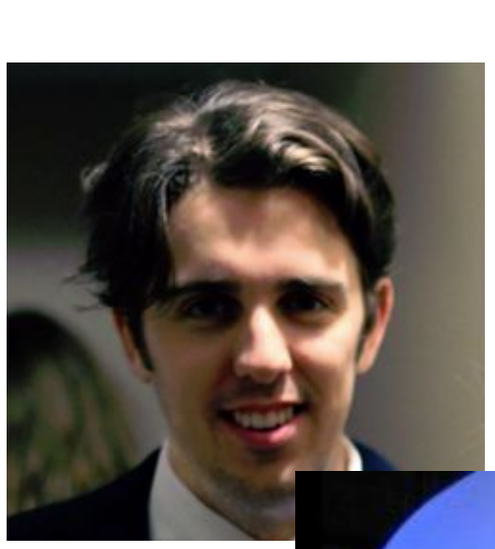
Abundance

Historians and other humanists can’t do it alone

We need collaboration!

Web Archives for Historical Research

Historians



Computer Scientists

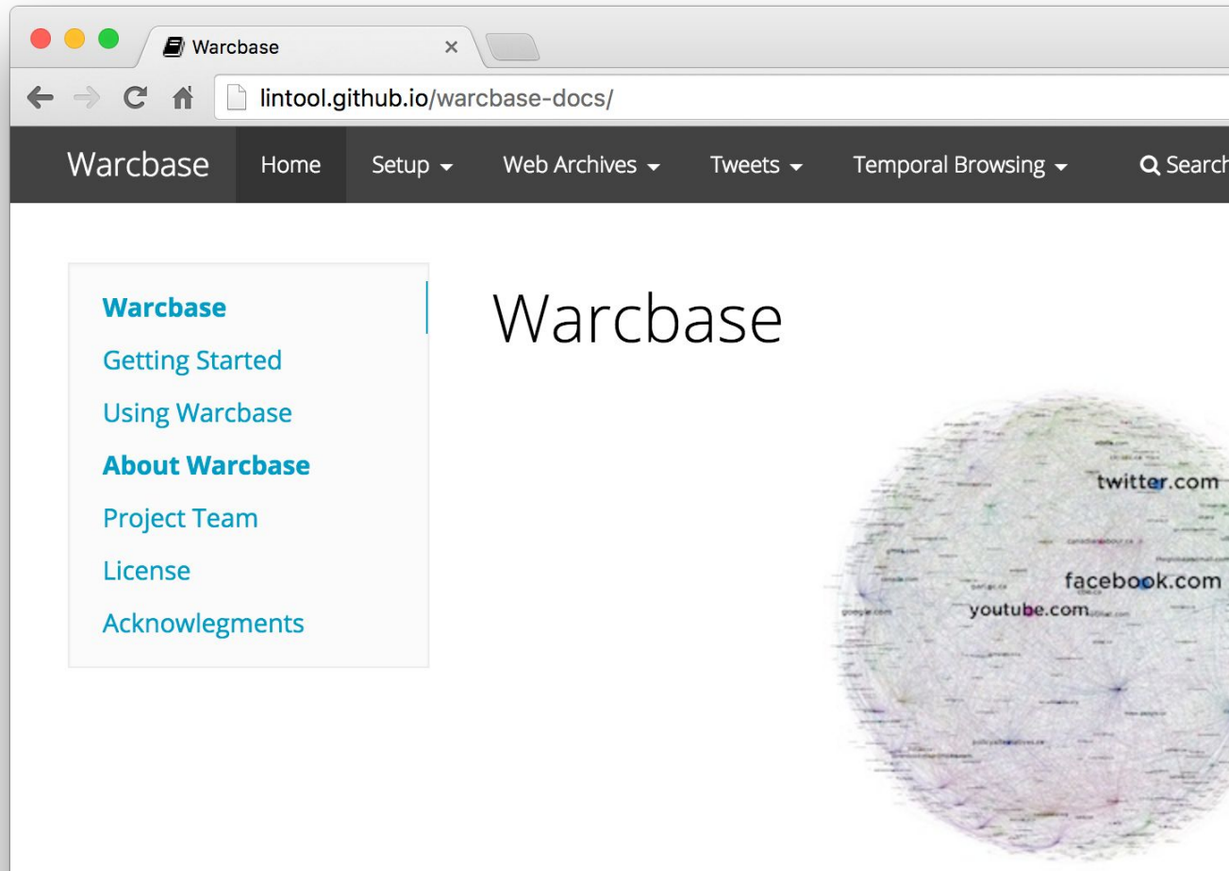
Librarians

Governance

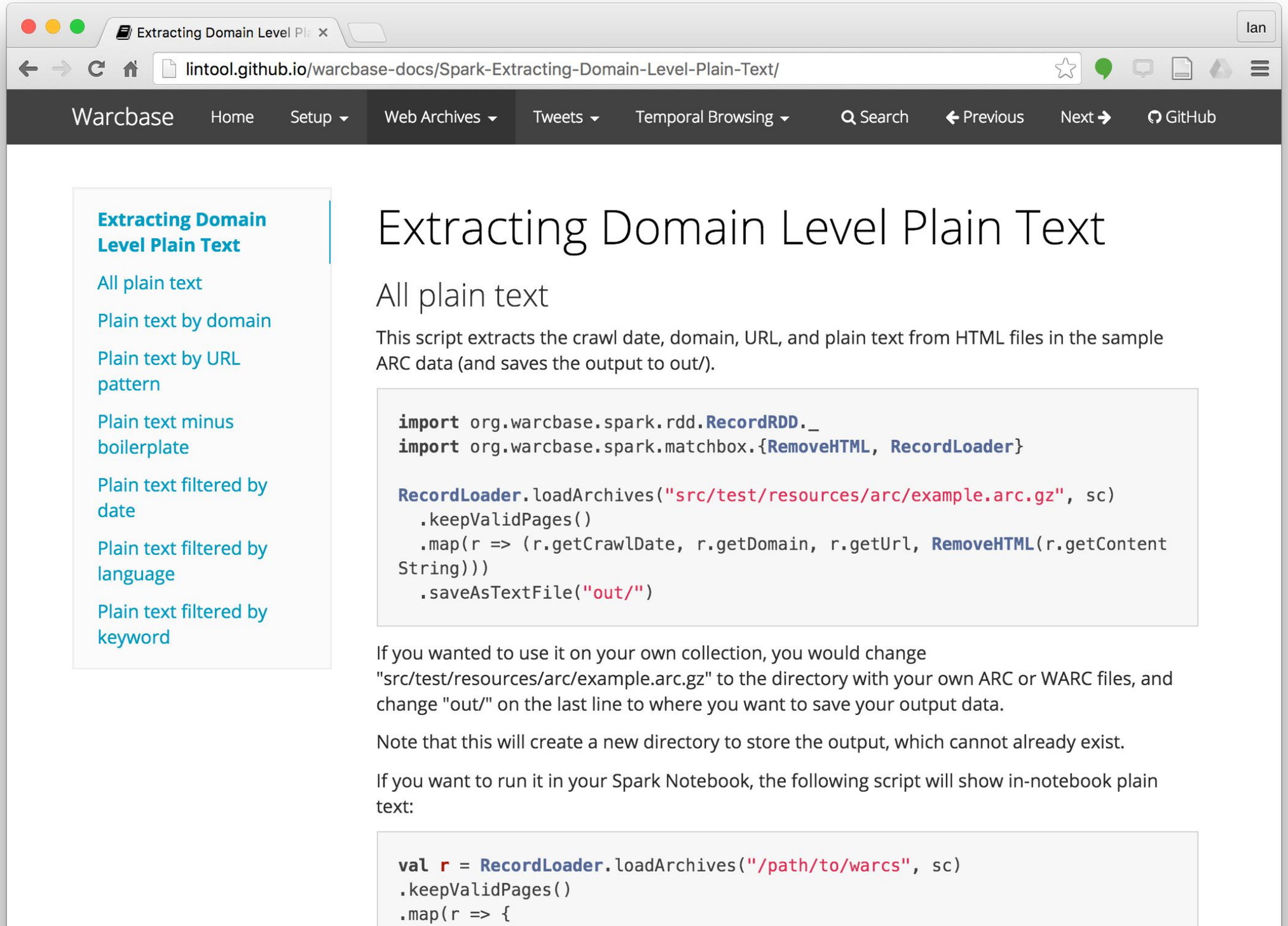
Warcbase

- **Jimmy Lin** (main developer, CS/lead), **Ian Milligan** (co-lead, history), **Jeremy Wiebe** (history/PhD), **Alice Zhou** (computer science, undergrad), **Youngbin Kim** (computer science, undergrad), **Nick Ruest** (librarian @ York)

- Currently using it on the **GeoCities** and **Canadian Politics** web archives, as well as WALK (61 Archive-It collections, 6 institutions)

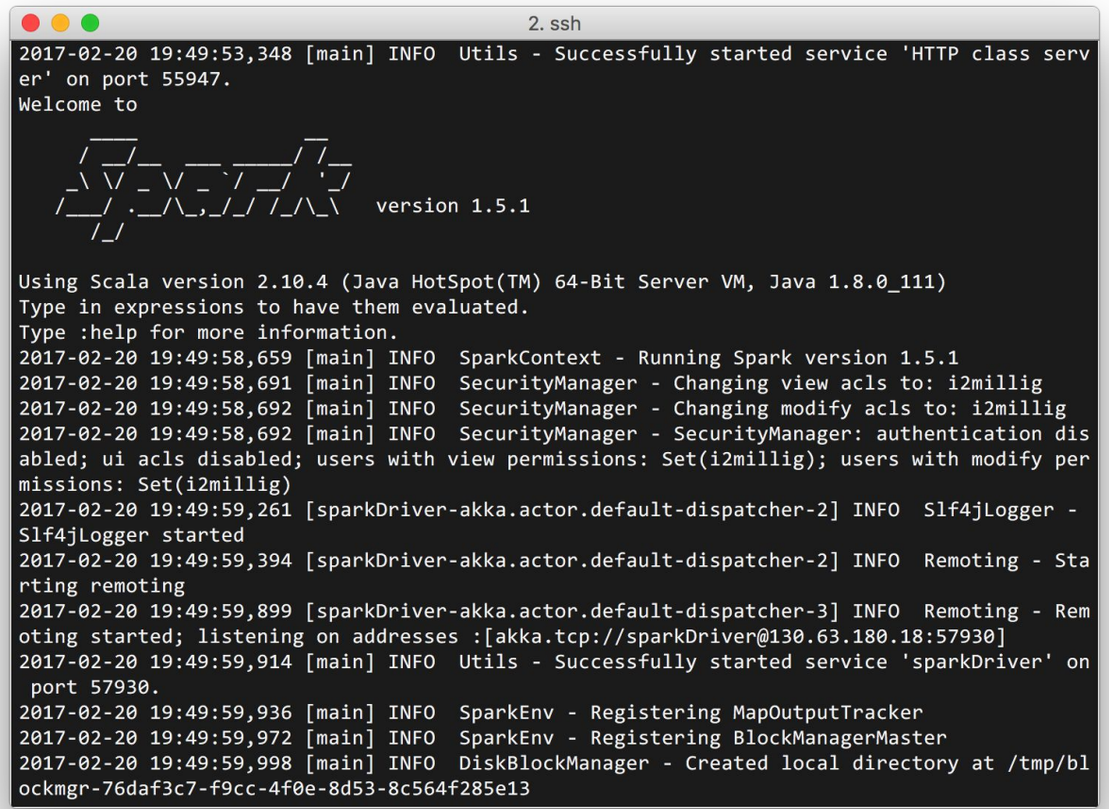


docs.warcbase.org



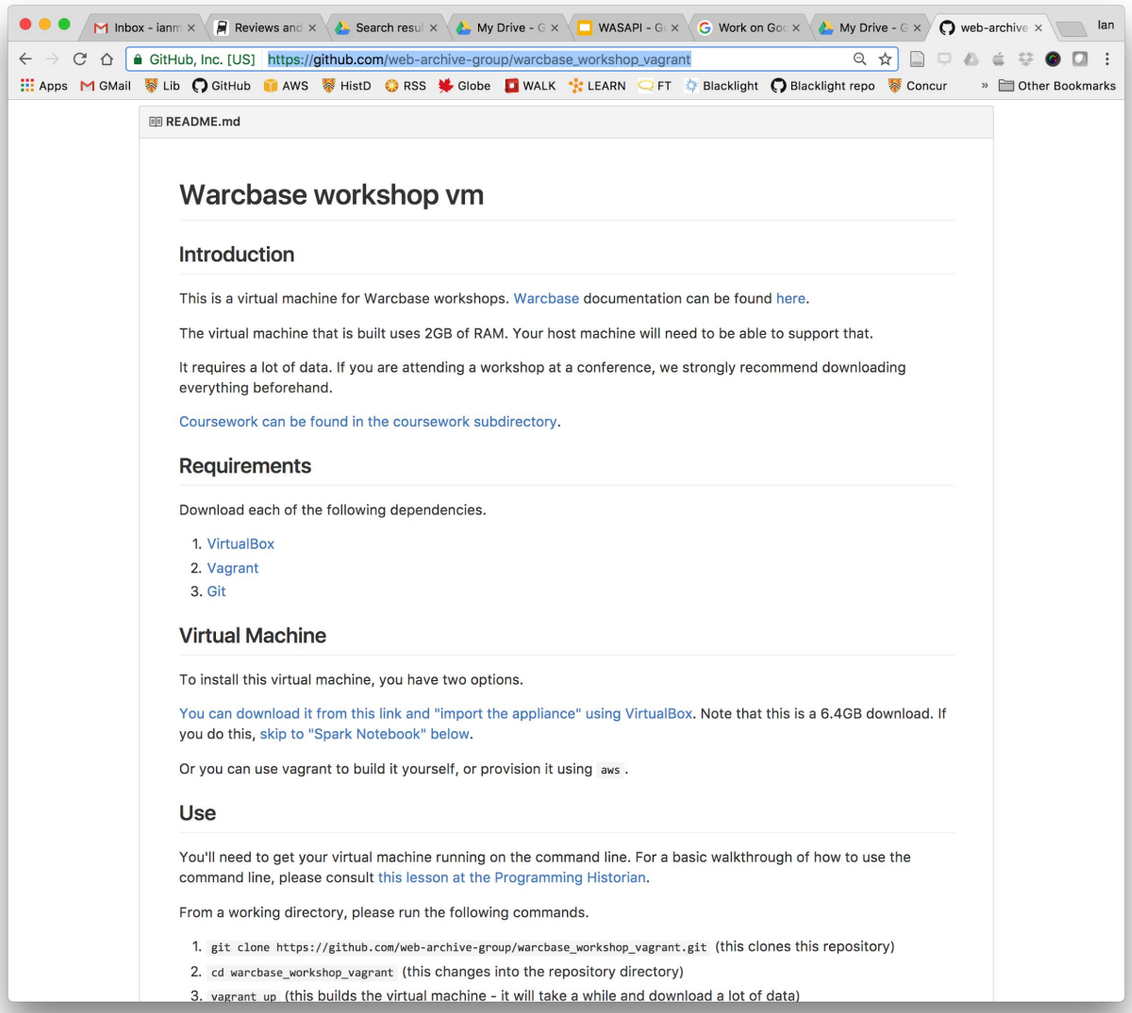
Installing Warcbase

- Install Spark, clone warcbase, build with maven
- Launch Spark Shell
- Run simple scripts (begin by copy-and-replacing our existing ones to get a feel for the interface)
- In the future, hoping to move to PySpark

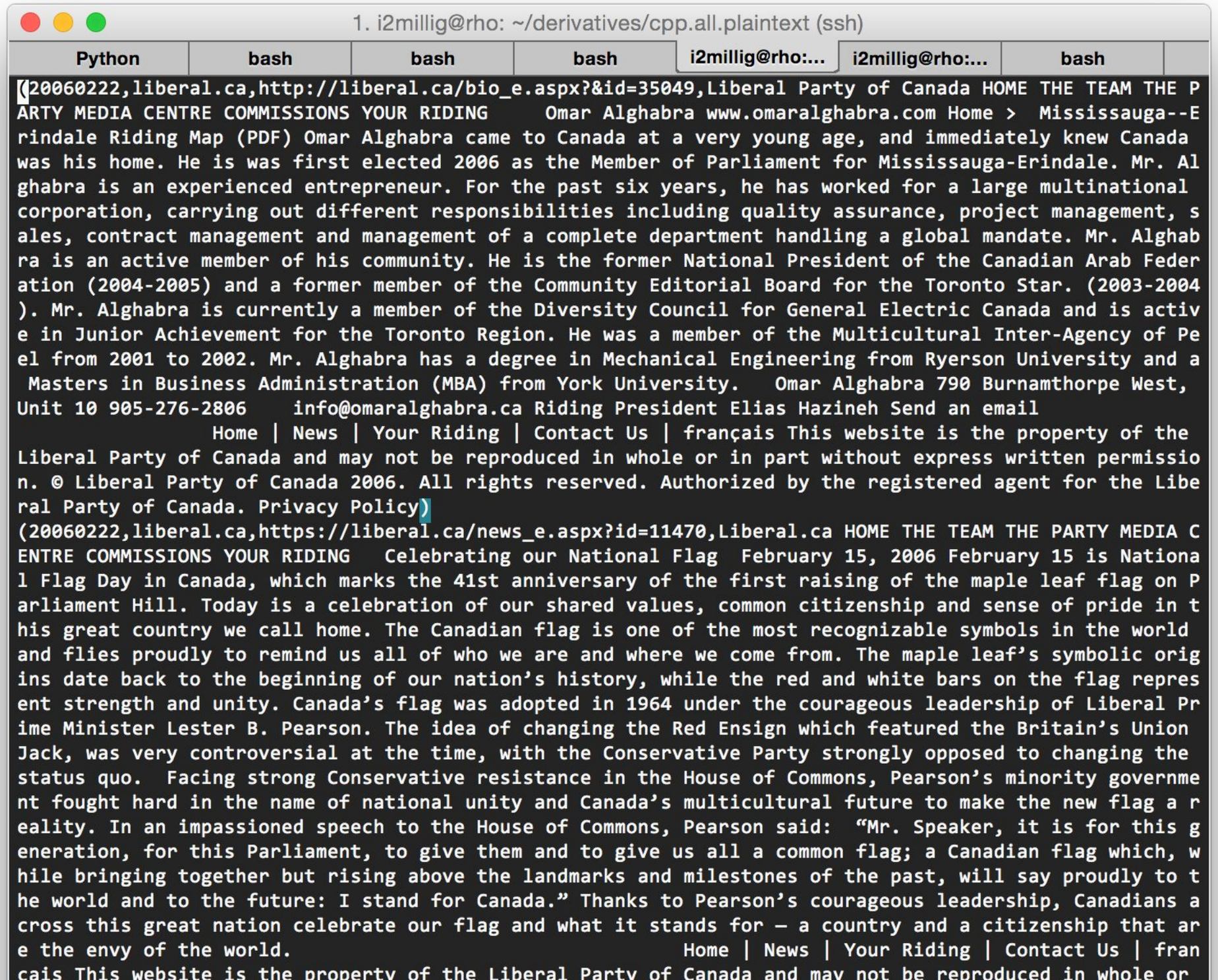


…. Or the Warcbase Vagrant machine!

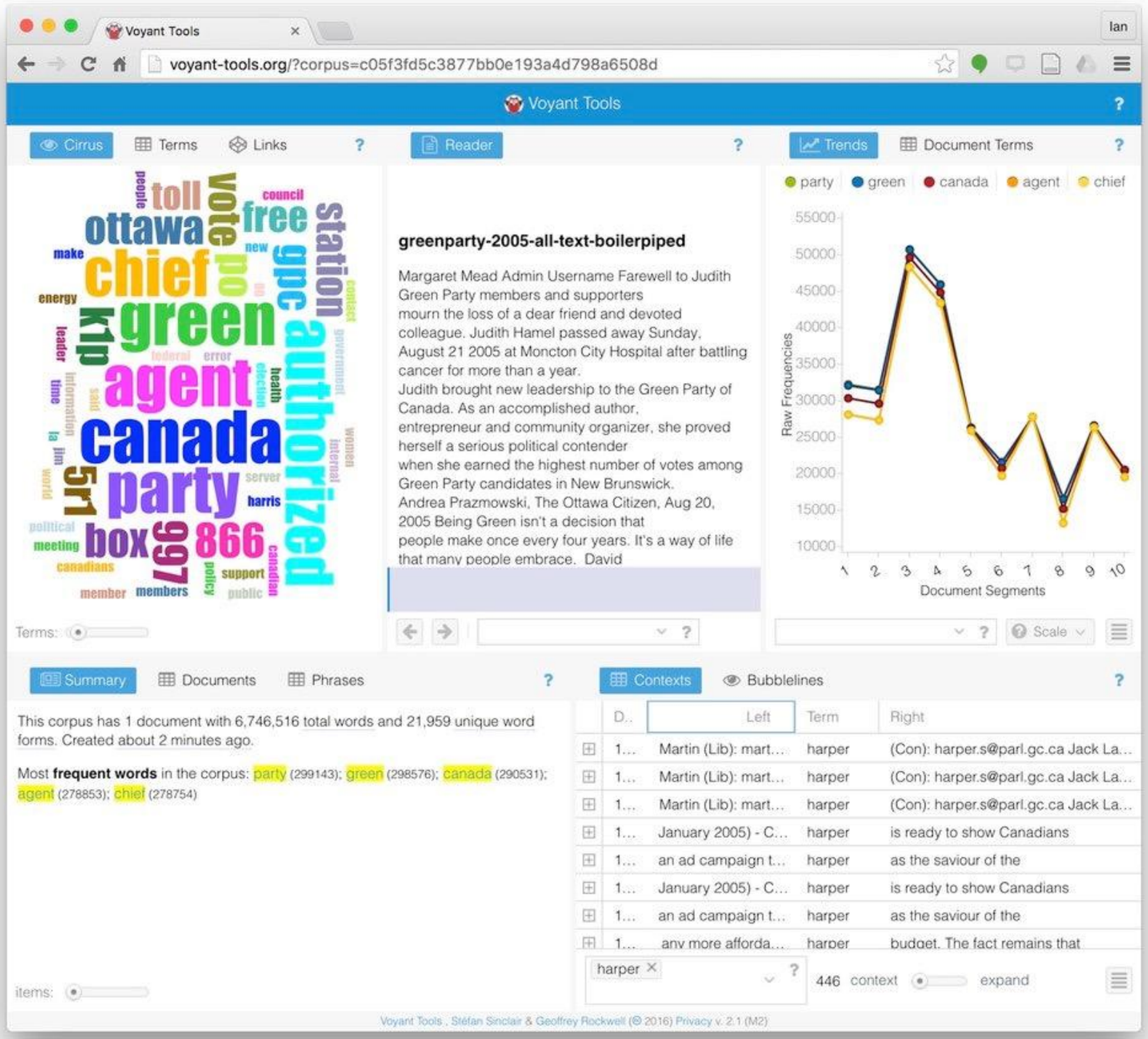
- Digital pedagogy FTW
- https://github.com/web-archive-group/warcbase_workshop_vagrant



Extract all Text



Extract all Text



Extract Entities

200606
Andrew Lewis
Bill
Bill Hulet
Brown
Bruce Abel
Bush
Camille Labchuk
Chandler
Cherfi
Chernushenko
David
David Chernsuhenko
David Chernushenko
David Kay
Derek Pinto
Ed Broadbent
Elizabeth May
Eric Walton
Fannon
Gomery
Green
Harper
Harris
Jim
Jim Fannon
Jim Harris
Jim Harris Speech
John
Julie Baribeau
Junker
Kevin Colton
Labchuk
Layton
Leonardo DiCaprio
Manley
Mark Brooks
Mark MacGillivray
Martin
Michael Robinson
Milliken
Paul Martin

200607
Adrianne Carr
Andrew Lewis
Bill
Bill Hulet
Brown
Bruce Abel
Bush
Camille Labchuk
Chandler
Cherfi
Chernushenko
David
David Chernsuhenko
David Chernushenko
David Chernushenko E...
David Kay
Derek Pinto
Dietrich
Ed Broadbent
Elizabeth May
Eric Walton
Fannon
Gomery
Green
Harper
Harris
Jim
Jim Fannon
Jim Harris
Jim Harris Speech
John
Julie Baribeau
Junker
Kevin Colton
Labchuk
Layton
Manley

200608
Adrianne Carr
Allan Gribbin
Amélie Gingras
Andrew Lewis
Bill
Bill Hulet
Brown
Bruce Abel
Bush
Chandler
Cherfi
Chernushenko
Clements Verhoeven
David
David Chernushenko
David Kay
Derek Pinto
Dietrich
Ed Broadbent
Elizabeth May
Eric Walton
Fannon
Gomery
Green
Harper
Harris
Jim
Jim Harris
Jim Harris Speech
John
Junker
Kevin Colton
Kootenay-Columbia Jo...
Labchuk
Lawrence Redfern
Layton
Manley

200609
Adrianne Carr
Amélie Gingras
Brown
Bruce Abel
Bush
Cameron Wigmore
Chandler
Cherfi
Chernushenko
Chretien
David
David Chernushenko
David Kay
Derek Pinto
Dietrich
Dion
Elizabeth
Elizabeth May
Elizabeth Peloza
Eric Walton
Gomery
Green
Harper
Harris
Jasper
Jim
Jim Harris
Jim Harris Speech
John
Labchuk
Lougheed
Mackenzie
Manley
Martin
May
Mona Elaine Adilman ...
Paul Martin
Peter Foster
Pierre Pettigrew
Schiller

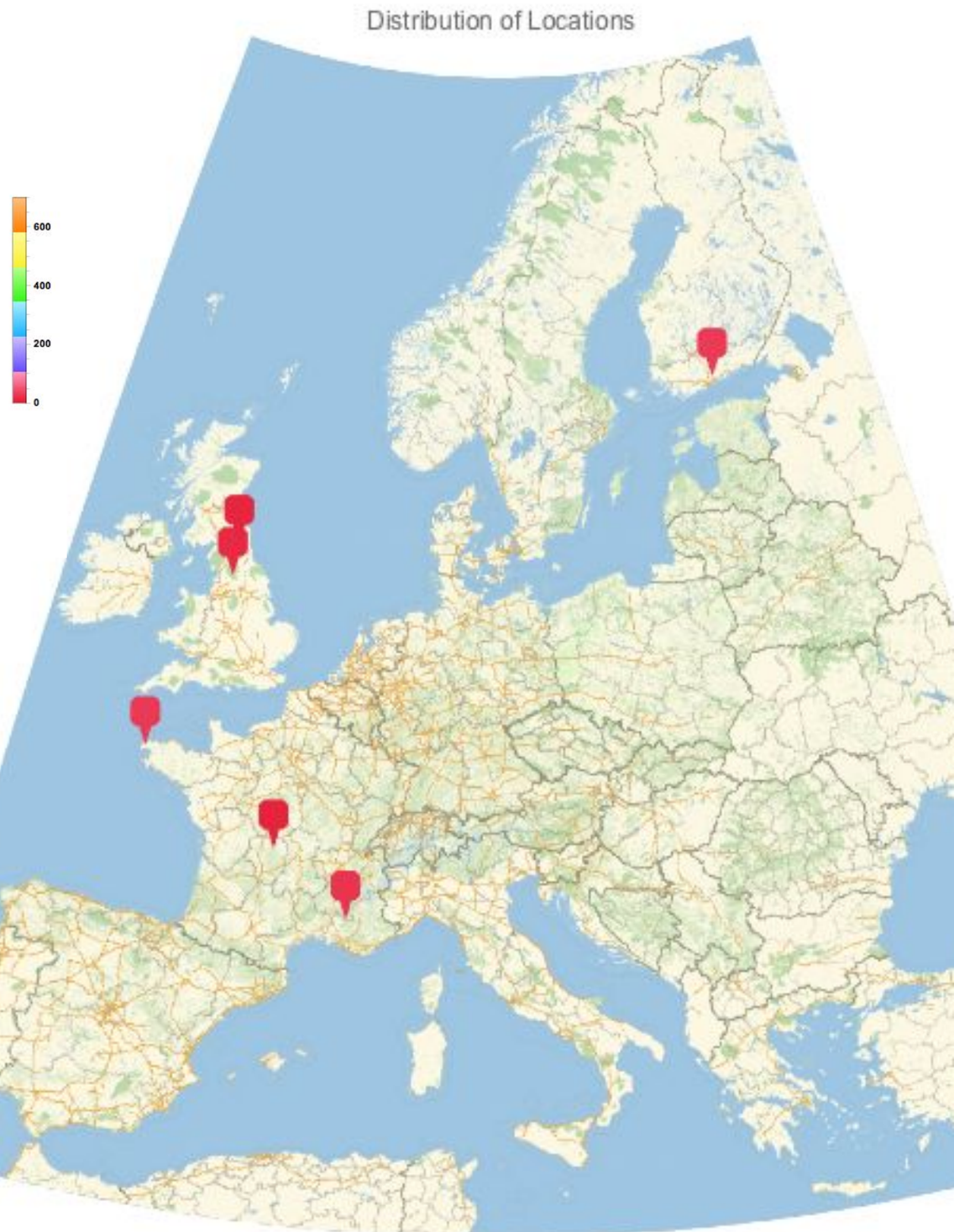
Elizabeth May
10 mentions

200610
Ambrose
Andrew Lewis
Bill
Bridget Doherty
Bush
Carol Gudz
Catharine Johannson
Chandler
Cherfi
Chernushenko
Daphne Wysham
David
David Chernushenko
David Kay
Derek
Derek Pinto
Dundas
Elizabeth
Elizabeth Goes
Elizabeth May
Elizabeth May Say
Eric Walton
Gagnon
Gomery
Green
Grenon
Halton
Harper
Harris
Jim
Jim Harris
John
Jude Larkin
Judith
Kyle Grice
Labchuk
Manley
Mark MacGillivray
Martin
May
Melanie Ransom
Michael Grayson
Michele
Paul Martin
Richard Reble
Sharon Labchuk

200611
Ambrose
Andrew Lewis
Bill
Bill Clinton
Bush
Chandler
Cherfi
Chernushenko
Chris Alders
Daphne Wysham
David
David Chernushenko
David Cox
David Kay
David Suzuki
Derek
Derek Pinto
Dundas
Edward Burtynsky
Elizabeth
Elizabeth May
Eric Walton
Garth Turner
Gomery
Green
Halton
Harper
Harris
Jim
Jim Harris
Jim Harris Speech
John
Julie Baribeau
Labchuk
Manley
Margaret
Mark MacGillivray
Martin
May
Paul
Paul Martin
Ross
Sharon Labchuk

Extract Entities

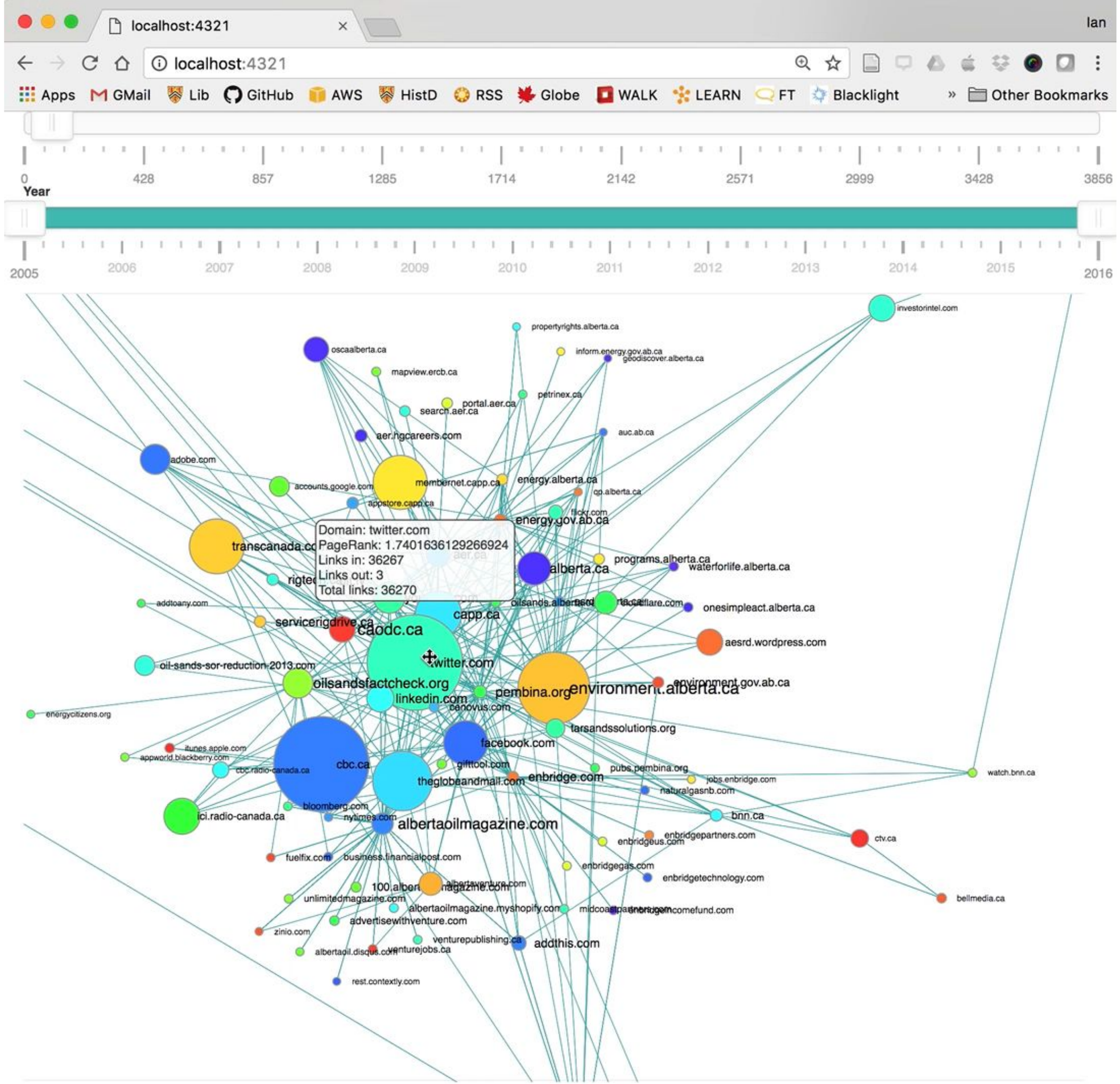
Extract Entities



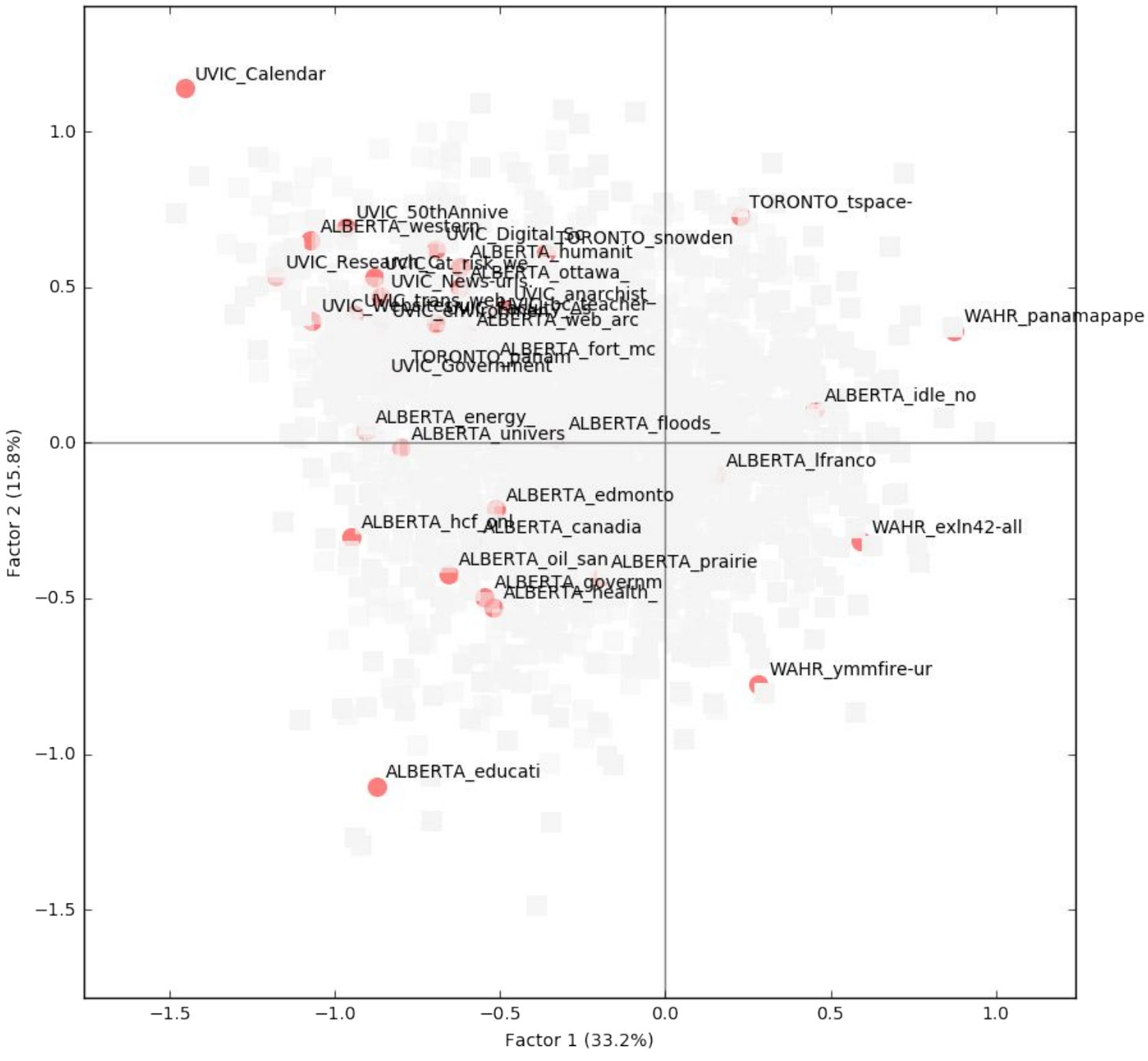

```
In[95]:= int = SemanticInterpretation[#] & /@ processedfreq[[All, 1]]
```

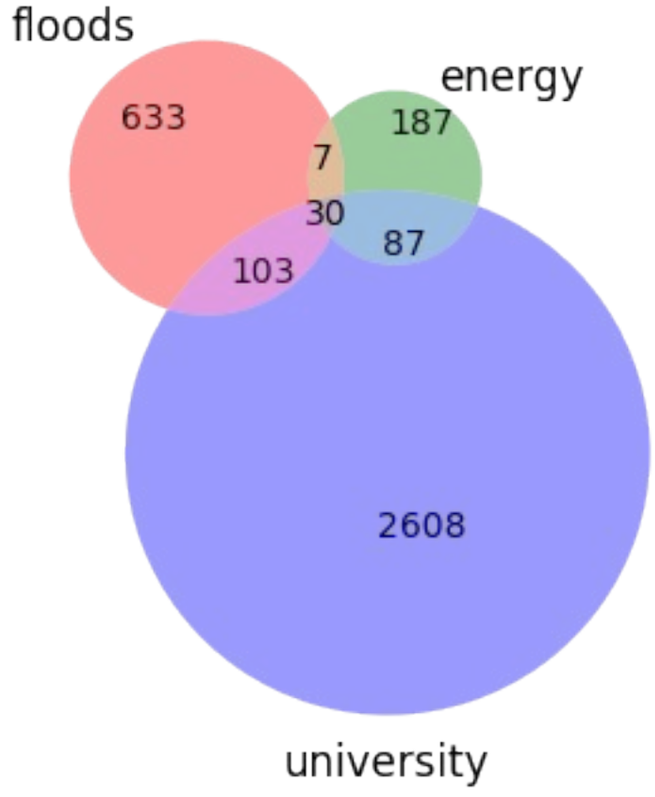
Out[95]= { Canada , Calgary , Colombia , Montreal , $Failed, Ontario, Canada , Afghanistan ,
Ottawa , Manitoba, Canada , British Columbia, Canada , Toronto , Nova Scotia, Canada ,
Saskatchewan, Canada , Quebec, Canada , Alberta, Canada , Winnipeg , Washington ,
Canada , Nunavut, Canada , White Horse , United States , Petawawa , Mumbai ,
Lima , The Americas , Saskatoon , Peru , Edmonton , Mexico , Chicoutimi-Jonquiere ,
New Brunswick, Canada , United States , Newfoundland and Labrador, Canada , Qandahar ,
$Failed, Asia-Pacific , Newfoundland and Labrador, Canada , Victoria , United States ,
Quebec City , India , $Failed, Atlantic Ocean , St. Germain , Durham , Battle of Waterloo ,

Finding Sites of Interest


localhost:4321
Ian
localhost:4321
Apps
GMail
Lib
GitHub
AWS
HistD
RSS
Globe
WALK
LEARN
FT
Blacklight
Other Bookmarks
0
428
857
1285
1714
2142
2571
2999
3428
3856
Year
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
investorintel.com
propertyrights.alberta.ca
oscaalberta.ca
inform.energy.gov.ab.ca
geodiscover.alberta.ca
mapview.ercb.ca
petrinex.ca
portal.aer.ca
search.aer.ca
aer.hgcareers.com
auc.ab.ca
adobe.com
accounts.google.com
membernet.capp.ca
energy.alberta.ca
cp.alberta.ca
appstore.capp.ca
flickr.com
energy.gov.ab.ca
Domain: twitter.com
PageRank: 1.7401636129266924
Links in: 36267
Links out: 3
Total links: 36270
transcanada.co
programs.alberta.ca
waterforlife.alberta.ca
alberta.ca
addtoany.com
capp.ca
onesimpleact.alberta.ca
servicerigdrive.ca
caodc.ca
aesrd.wordpress.com
twitter.com
oil-sands-sor-reduction-2013.com
oilsandsfactcheck.org
environment.gov.ab.ca
pembina.org
linkedin.com
energycitizens.org
tarsandssolutions.org
facebook.com
itunes.apple.com
appworld.blackberry.com
cbc.radio-canada.ca
cbc.ca
gifttool.com
pubs.pembina.org
watch.bnn.ca
theglobeandmail.com
enbridge.com
jobs.enbridge.com
naturalgasnb.com
bloomberg.com
ici.radio-canada.ca
nytimes.com
albertaoilmagazine.com
bnn.ca
enbridgepartners.com
ctv.ca
enbridgeus.com
fuelfix.com
business.financialpost.com
enbridgegas.com
enbridgetechnology.com
albertaventure.com
unlimitedmagazine.com
bellmedia.ca
albertaoilmagazine.myshopify.com
advertisewithventure.com
zinio.com
venturepublishing.ca
addthis.com
albertaoil.disqus.com
venturejobs.ca
rest.contextly.com

Exploring collection coverage and curatorial models





So we have this tool... how have we used it?

Scaling up

Web Archives for Longitudinal Knowledge

- **Ian Milligan** (Co-PI, UW) + **Nick Ruest** (Co-PI, York), w/ **Geoff Harder**, **Todd Suomela**, **Sonya Betz**, **Peter Binkley**, **Geoffrey Rockwell, Umar Qasim** (Alberta), **Jefferson Bailey** (Internet Archive), and **John Simpson** (Compute Canada).

Web Archives for Longitudinal Knowledge

- 6 universities (SFU, Toronto, Alberta, Victoria, Winnipeg, and Dalhousie)

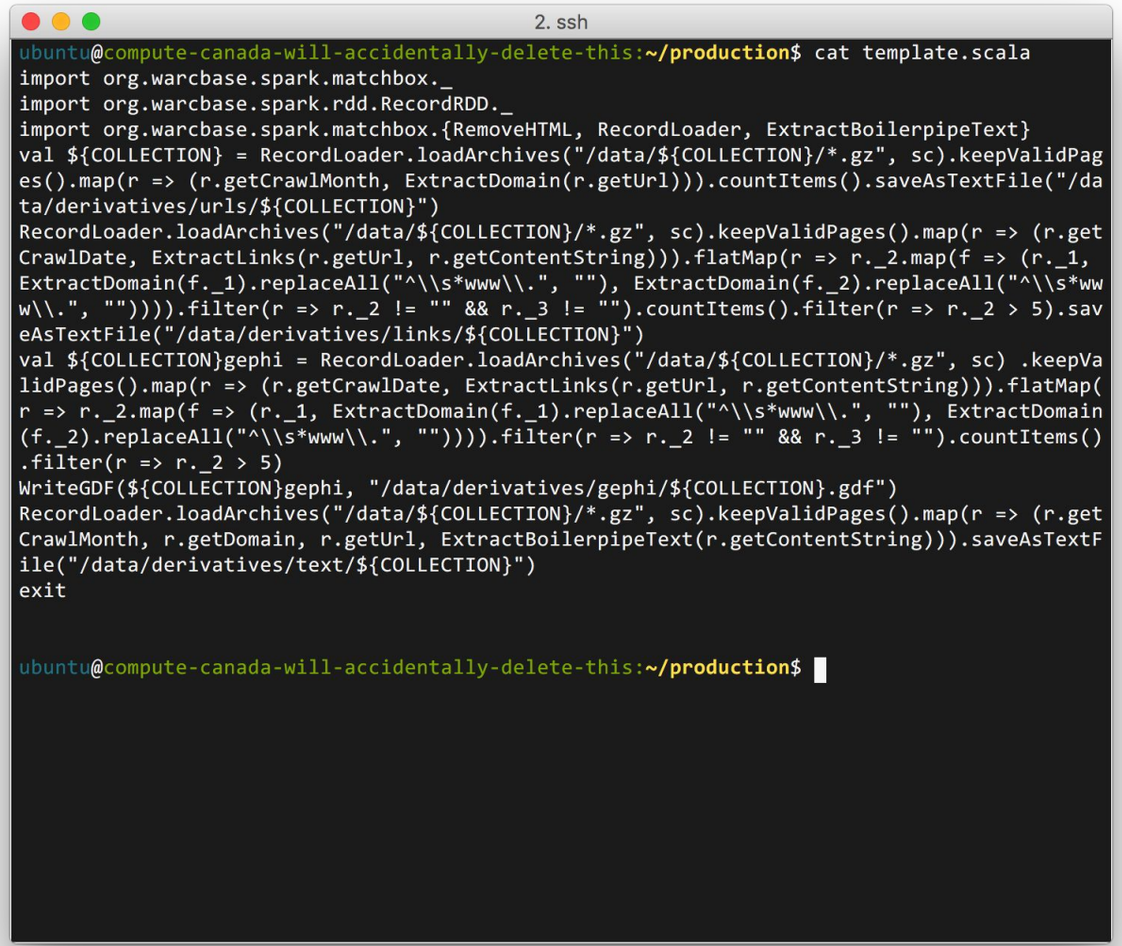
- 61 collections

- 16 TB of web archival collections

- Ingesting them all to generate warcbase derivatives and searchable index

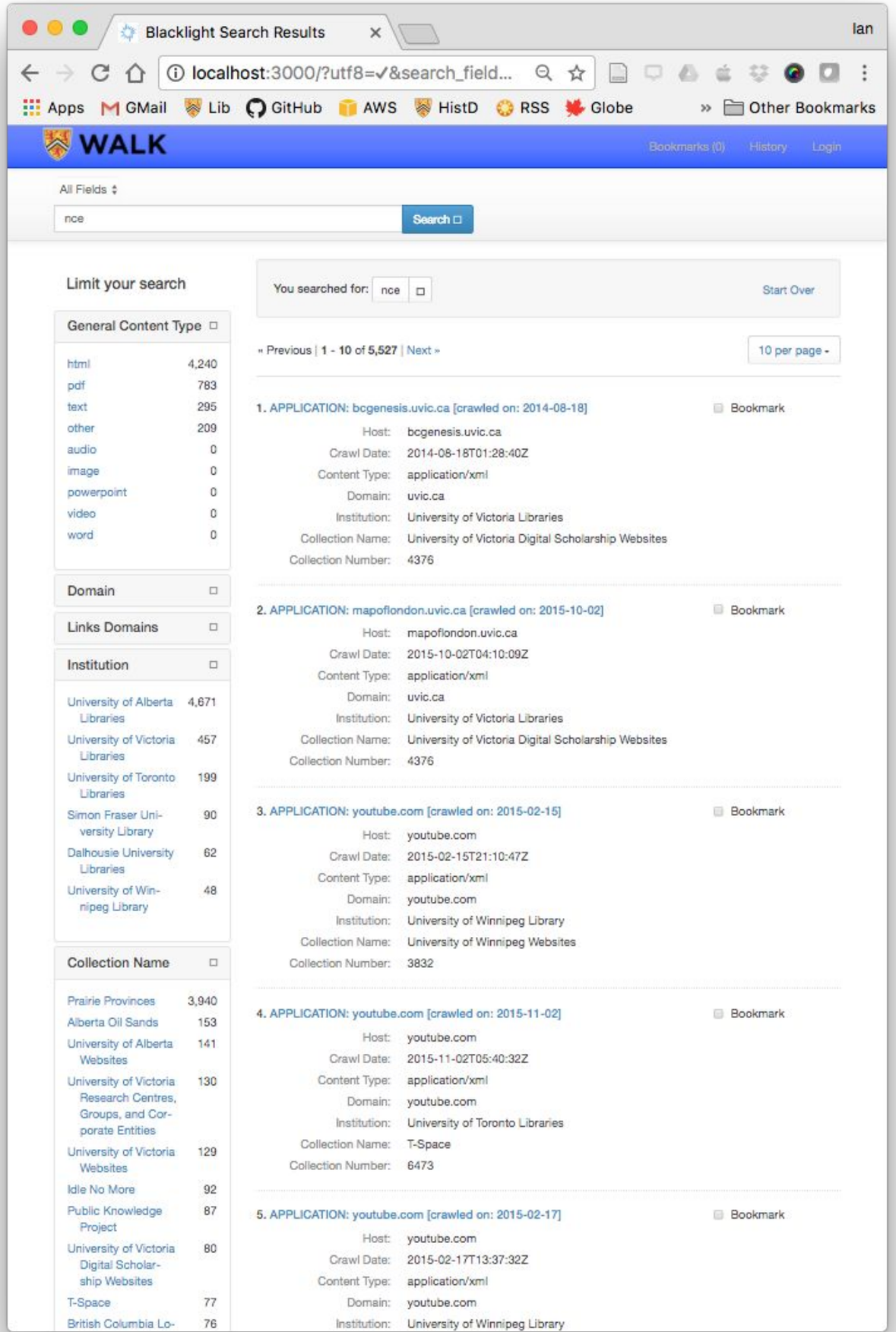
Web Archives for Longitudinal Knowledge

- Mass Ingest Script to take each collection and:
  - Extract hyperlinks and generate gephi files;
  - Extract all URLs and domain counts;
  - Extract plain text;



WebArchives.ca (Blacklight)

- Warcbase with a GUI front end
- Archivists/Librarians bring their collections, generate derivatives
- Hosted with Compute Canada



Where are we at now?

Getting the data;

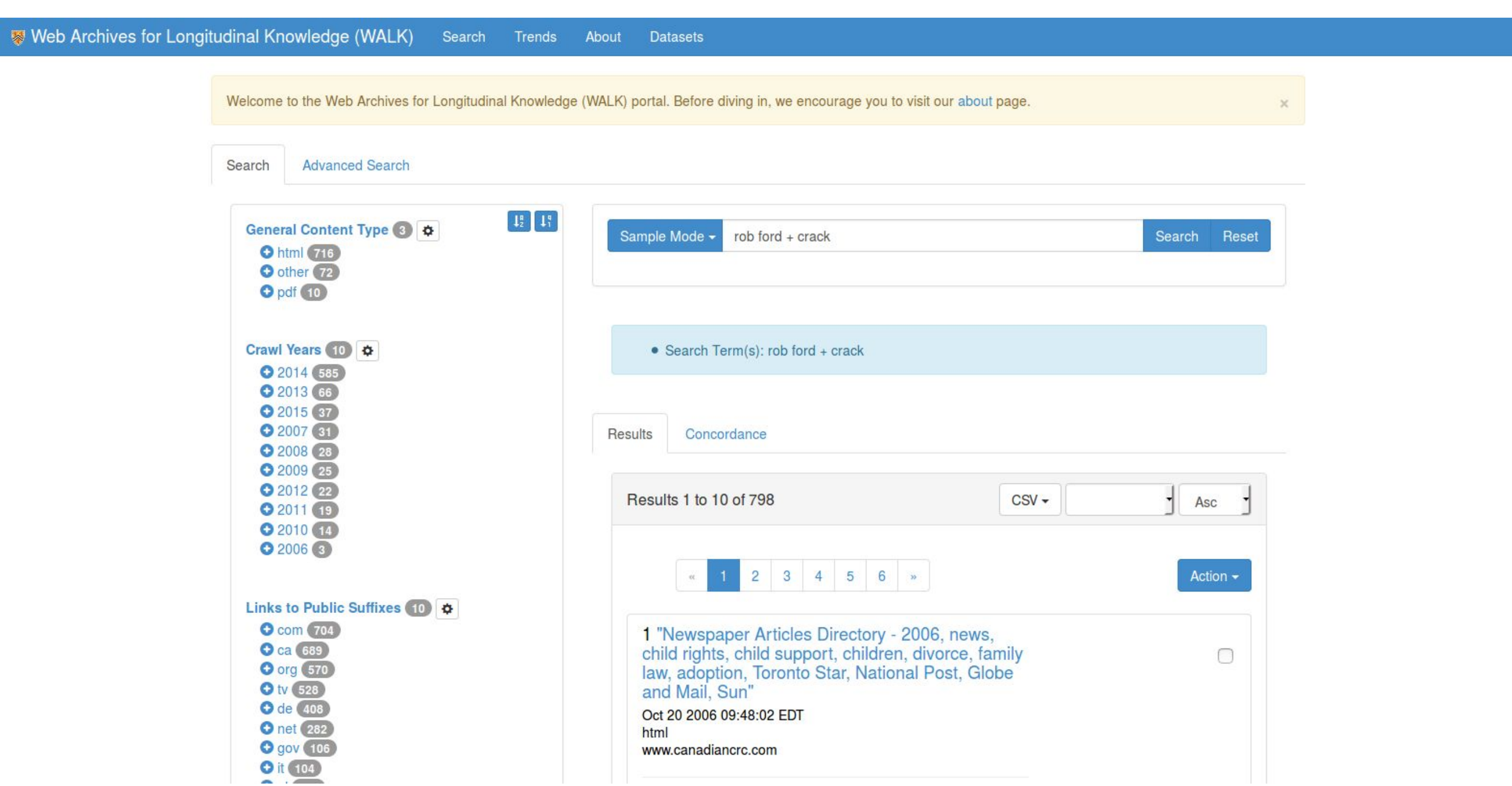
Definitely a need for a standardized interaction model!

UKWA
and
Andy Jackson,
Thank you!

Indexing

```
2017-02-21 21:46:27 ERROR HTMLAnalyser:134 - Failed to canonicalise host: French Language Summer Camps Level I: org.apache.commons.httpclient.URIException: gnu.inet.encoding.IDNAException: Contains
non-LDH characters. french%20language%20summer%20camps%20level%20i
2017-02-21 22:07:45 ERROR HTMLAnalyser:134 - Failed to canonicalise host: josephmoreau.schoolappointments.
com: org.apache.commons.httpclient.URIException: gnu.inet.encoding.IDNAException: Contains non-LDH characters. josephmoreau.schoolappointments.%20com
2017-02-21 22:07:52 ERROR HTMLAnalyser:134 - Failed to canonicalise host: josep
hmoreau.schoolappointments.com: org.apache.commons.httpclient.URIException: gnu.inet.encoding.IDNAException: Contains non-LDH characters. josep%20hmoreau.schoolappointments.com
2017-02-21 22:21:28 INFO  Instrument:155 - Performance statistics
WARCIndexerCommand.main#total(#=0, time=0.00ms, avg=0.00#/ms 0.00ms/#, 0.00%)
  WARCIndexerCommand.parseWarcFiles#docdelivery(#=2089123, time=1706007.36ms, avg=1.22#/ms 0.82ms/#, 2.07%)
    WARCIndexerCommanc.checkSubmission#solradd(#=41782, time=1701567.49ms, avg=0.02#/ms 40.72ms/#, 2.06%)
  WARCIndexerCommand.parseWarcFiles#startup(#=1, time=3372.23ms, avg=0.00#/ms 3372.23ms/#, 0.00%)
  WARCIndexerCommand.commit#success(#=30, time=2663523.35ms, avg=0.00#/ms 88784.11ms/#, 3.23%)
  WARCIndexerCommand.parseWarcFiles#fullarcprocess(#=30, time=1445718780.22ms, avg=0.00#/ms 48190626.01ms/#, 1754.28%)
  WARCIndexerCommand.parseWarcFiles#solrdocCreation(#=8651855, time=77583908.73ms, avg=0.11#/ms 8.97ms/#, 94.14%)
    WARCIndexer.extract#total(#=1740156, time=74867498.43ms, avg=0.02#/ms 43.02ms/#, 90.85%)
      TextAnalyzers#total(#=1740156, time=5724356.71ms, avg=0.30#/ms 3.29ms/#, 6.95%)
        PostcodeAnalyzer(#=1701619, time=46774.75ms, avg=36.38#/ms 0.03ms/#, 0.06%)
        LanguageAnalyzer#total(#=1701619, time=5334955.20ms, avg=0.32#/ms 3.14ms/#, 6.47%)
          LanguageDetector#startup(#=1, time=385.84ms, avg=0.00#/ms 385.84ms/#, 0.00%)
          LanguageDetector.detectLanguage#li(#=1701619, time=2859452.22ms, avg=0.60#/ms 1.68ms/#, 3.47%)
            LanguageIdentifier.addProfile(#=28, time=29.13ms, avg=0.96#/ms 1.04ms/#, 0.00%)
            LanguageIdentifier#matchlanguageprofile(#=1701619, time=2551253.69ms, avg=0.67#/ms 1.50ms/#, 3.10%)
              LanguageProfile.distanceInterleaved#total(#=47645332, time=2539920.08ms, avg=18.76#/ms 0.05ms/#, 3.08%)
                LanguageProfile.Interleaved.update(#=1701647, time=216333.19ms, avg=7.87#/ms 0.13ms/#, 0.26%)
                LanguageProfile.distanceInterleaved#dist(#=47645332, time=2236326.36ms, avg=21.31#/ms 0.05ms/#, 2.71%)
            LanguageProfile#profilewriter(#=1701619, time=306254.26ms, avg=5.56#/ms 0.18ms/#, 0.37%)
          LanguageDetector.detectLanguage#ld(#=1701562, time=2466748.82ms, avg=0.69#/ms 1.45ms/#, 2.99%)
        FuzzyHashAnalyzer(#=1701619, time=337414.28ms, avg=5.04#/ms 0.20ms/#, 0.41%)
      WARCIndexer.extract#hashstreamwrap(#=2089123, time=3746222.06ms, avg=0.56#/ms 1.79ms/#, 4.55%)
      WARCIndexer.extract#analyzetikainput(#=1740156, time=64287494.76ms, avg=0.03#/ms 36.94ms/#, 78.01%)
        WARCPayloadAnalyzers.analyze#total(#=1740156, time=64284418.35ms, avg=0.03#/ms 36.94ms/#, 78.00%)
          WARCPayloadAnalyzers.analyze#firstbytes(#=1740156, time=17903.05ms, avg=97.20#/ms 0.01ms/#, 0.02%)
          ARCNameAnalyzer.analyze(#=1740156, time=13765.56ms, avg=126.41#/ms 0.01ms/#, 0.02%)
          XMLAnalyzer.analyze(#=7975, time=7579.49ms, avg=1.05#/ms 0.95ms/#, 0.01%)
          WARCPayloadAnalyzers.analyze#arcname(#=1740156, time=14751.68ms, avg=117.96#/ms 0.01ms/#, 0.02%)
          WARCPayloadAnalyzers.analyze#tikasolrextract(#=1740156, time=33048823.86ms, avg=0.05#/ms 18.99ms/#, 40.10%)
            TikaExtractor.extract#detect(#=1740156, time=8280514.40ms, avg=0.21#/ms 4.76ms/#, 10.05%)
            TikaExtractor.extract#extract(#=1735438, time=616945.84ms, avg=2.81#/ms 0.36ms/#, 0.75%)
            TikaExtractor.extract#parse(#=1735438, time=23454988.10ms, avg=0.07#/ms 13.52ms/#, 28.46%)
          ImageAnalyzer.analyze#facesanddominant(#=2547, time=1933767.38ms, avg=0.00#/ms 759.23ms/#, 2.35%)
          WARCPayloadAnalyzers.analyze#droid(#=1740156, time=10547860.60ms, avg=0.16#/ms 6.06ms/#, 12.80%)
          HTMLAnalyzer.analyze#total(#=1683855, time=18072449.77ms, avg=0.09#/ms 10.73ms/#, 21.93%)
            HTMLAnalyzer.analyze#parser(#=1683855, time=12592642.58ms, avg=0.13#/ms 7.48ms/#, 15.28%)
              HtmlFeatureParser.parse#jsoupparse(#=1683855, time=9542197.47ms, avg=0.18#/ms 5.67ms/#, 11.58%)
              HtmlFeatureParser.parse#featureextract(#=1683855, time=2142947.86ms, avg=0.79#/ms 1.27ms/#, 2.60%)
          PDFAnalyzer.analyze(#=5926, time=636542.70ms, avg=0.01#/ms 107.42ms/#, 0.77%)
      WARCIndexer.extract#archeaders(#=2882905, time=693994.36ms, avg=4.15#/ms 0.24ms/#, 0.84%)
SolrRecord.removeControlCharacters#total(#=235107394, time=1551471.12ms, avg=151.54#/ms 0.01ms/#, 1.88%)
  SolrRecord.sanitiseUTF8(#=235107394, time=469633.81ms, avg=500.62#/ms 0.00ms/#, 0.57%)
Parsing Archive File [31/131]:/data/ALBERTA_lfrancophonie_de_louest_canadien/ARCHIVEIT-3835-SEMIANNUAL-22840-20141107010041205-00007-wbgrp-crawl104.us.archive.org-6446.warc.gz
```

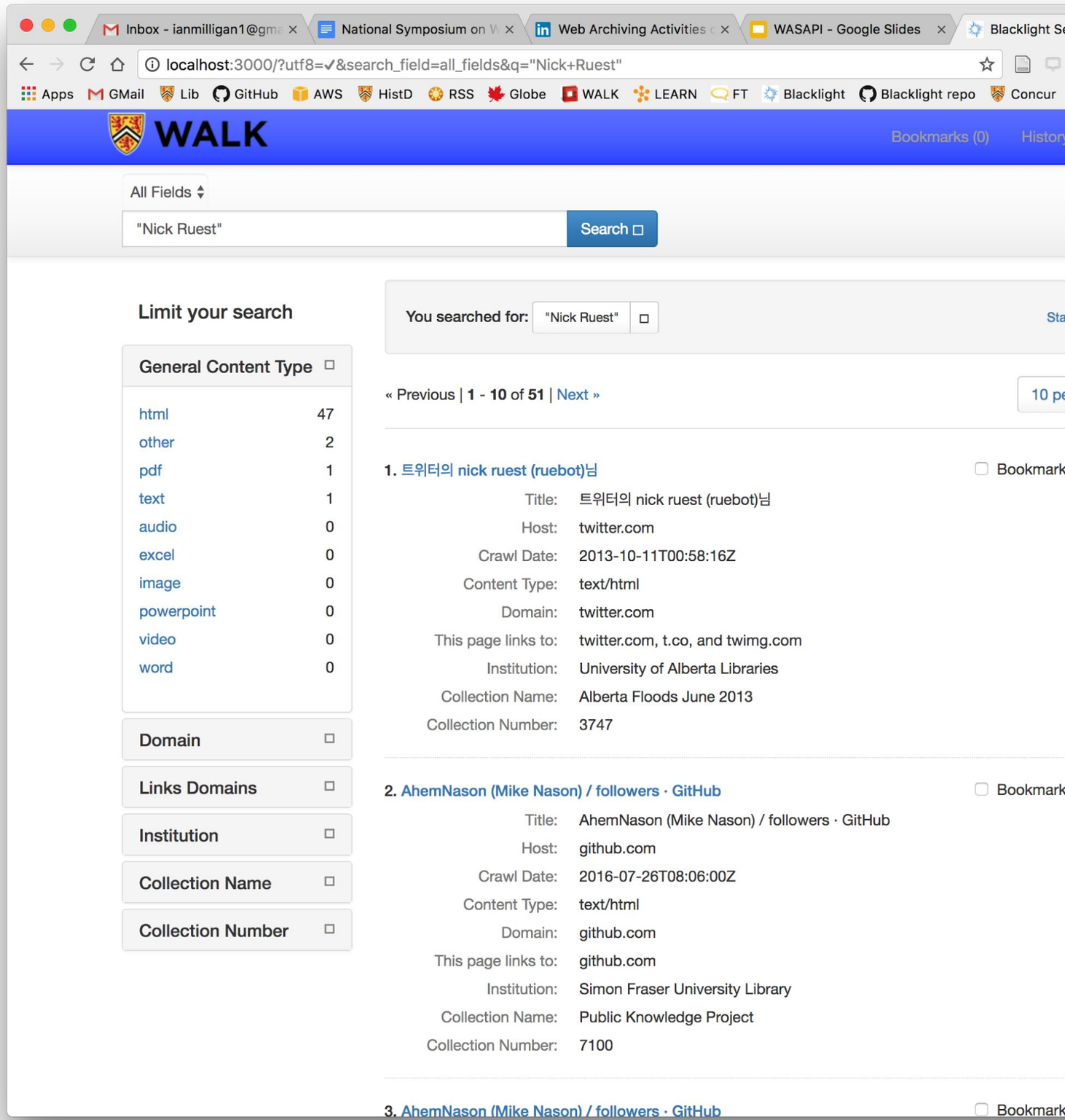
100 Million Solr Docs
525GB Solr index


Web Archives for Longitudinal Knowledge (WALK)
Search
Trends
About
Datasets
Welcome to the Web Archives for Longitudinal Knowledge (WALK) portal. Before diving in, we encourage you to visit our about page.
Search
Advanced Search
General Content Type 3
html 716
other 72
pdf 10
Crawl Years 10
2014 585
2013 66
2015 37
2007 31
2008 28
2009 25
2012 22
2011 19
2010 14
2006 3
Links to Public Suffixes 10
com 704
ca 689
org 570
tv 528
de 408
net 282
gov 106
it 104
Sample Mode
rob ford + crack
Search
Reset
Search Term(s): rob ford + crack
Results
Concordance
Results 1 to 10 of 798
CSV
Asc
« 1 2 3 4 5 6 »
Action
1 "Newspaper Articles Directory - 2006, news, child rights, child support, children, divorce, family law, adoption, Toronto Star, National Post, Globe and Mail, Sun"
Oct 20 2006 09:48:02 EDT
html
www.canadiancrc.com

Shine

WARClight

prototype



Finding famous people in WARClight, by Ian Millign

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Ontario

compute canada | calcul canada

UNIVERSITY OF WATERLOO

Thanks very much!

Ian Milligan
Assistant Professor
@ianmilligan1

Nick Ruest
Digital Assets Librarian
@ruebot

UNIVERSITY OF WATERLOO
FACULTY OF ARTS
Department of History

YORK UNIVERSITÉ UNIVERSITY